DOCUMENT RESUME

ED 174 636 TM 009 004

AUTHOR Leitner, Dennis W.
TITLE Using Multiple Regression to Interpret Chi-Square Contingency Table Analysis.
PUB DATE Apr 79
NOTE 9p.; Paper presented at the Annual Meeting of the American Educational Research Asscciation (63rd, San Francisco, Califcrria, April 8-12, 1979)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTCRS *Expectancy Tables; Hypothesis Testing; *Multiple Regression Analysis; *Nonparametric Statistics; Statistical Analysis
IDENTIFIERS *Chi Square

ABSTRACT

Statistics such as chi-square, phi, and Cramer's V are related to the R squared statistic of regression analysis. It is shown that the proportion of variance accounted fcr can be computed frcm mary contingency table situations. (JKS)

USING MULTIPLE REGRESSION TO INTERPRET CHI-SQUARE CONTINGENCY TABLE ANALYSIS

Dennis W. Leitner

Southern Illinois University at Carbondale

A paper presented at the annual meeting of the

American Educational Research Association

San Francisco, CA April, 1979

Session 24.17

A reinactment

Student: I have found it! A significant relationship between sex and political party affiliation. Look at this 2 x 2 contingency table and the chi-square value is significant at the .01 level.

Table 1. Frequencies of political party affiliation by sex of respondent.

	Democrat	Republican	
Male	60	40	$\chi_1^2 = 8.0$
Female	40	60	

Faculty Member (while scratching on a pad of paper and punching on his calculator): But are you sure that you have found a meaningful relationship?

S: What do you mean? It's significant and chi-squares with one degree of freedom seldom get beyond 6.

F (while fumbling through a file cabinet to get a scatter diagram): But what if I show you a bivariate scatterplot illustrating the strength of the relationship you have shown. (See Figure 1.)

S (crestfallen): But there is no relationship there.

F: Yes there is. A significant one at the .05 level of significance, for the 200 points, r = .2.

S: But that means only 4% of the variance is accounted for, 96% is unexplained.

F: Yes, and that is the strength of the relationship you have found with the chi-square analysis.

(The scene continues and ends with the student and faculty member commiserating at a local hangout.)

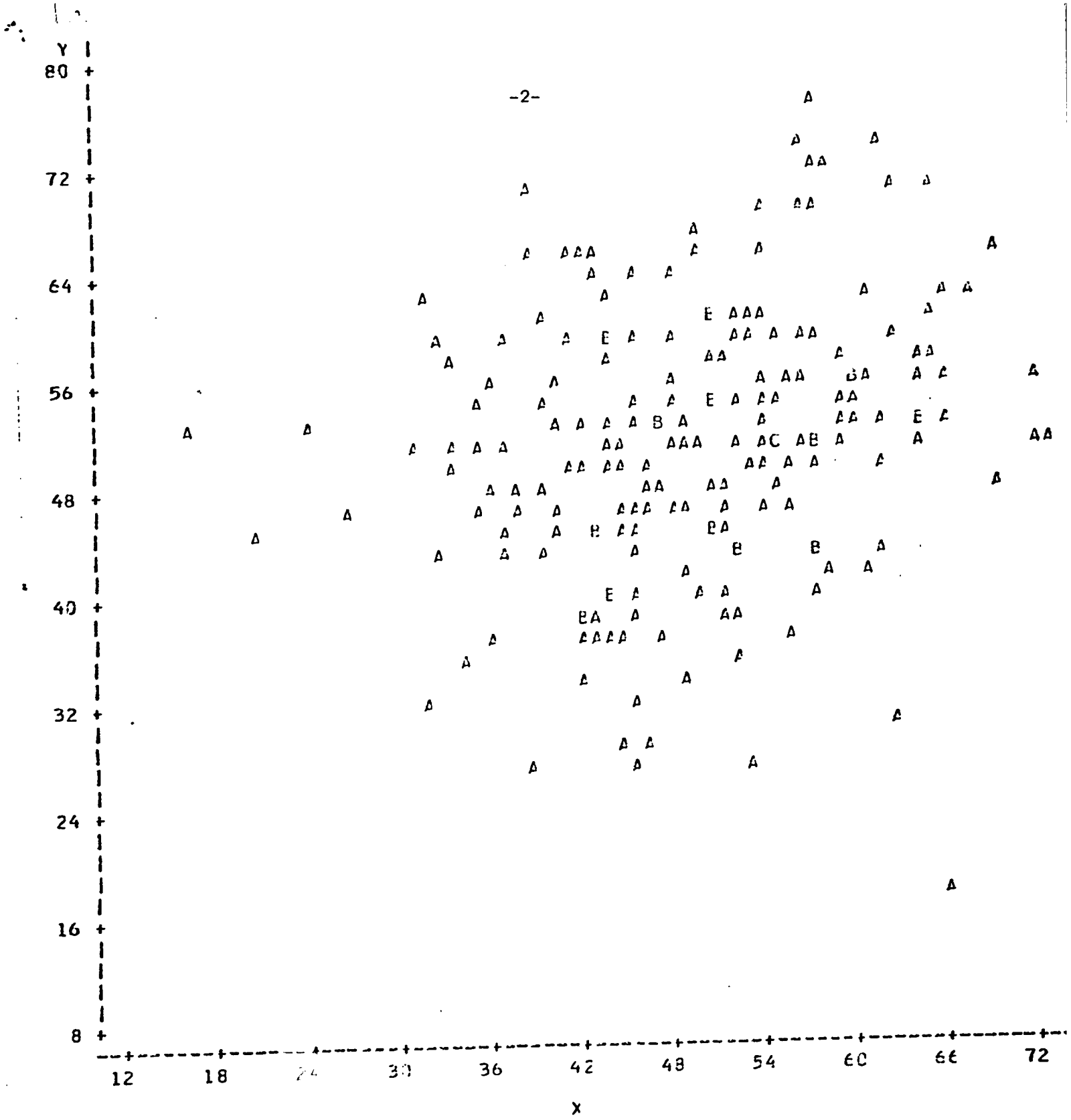

Figure 1. A scatterdiagram variables X and Y where $\mu_x = \mu_y = 50$, $\sigma_x = \sigma_y = 10$, N = 200, r = .2

The purpose of this paper is to emphasize the interpretation of R^2 in multiple regression that carries over to chi-square contingency table analysis.

2 x 2 Contingency Tables

Multiple regression aficionados would not have found themselves in the role of the student in the above scenario. By "dummy coding" sex and political party affiliation, and regression one of the other, an R^2 (actually r^2) value is obtained which is numerically equal to χ^2/N, where N is the number of subjects on which the two variables are measured (McNeil, Kelly, McNeil, 1975, pp. 246-248). In general, if row variable A has two levels and column variable B has two levels, let

$$X = \begin{cases} 1 & \text{if observation is from level 1 of A} \\ 0 & \text{otherwise} \end{cases} \qquad (1)$$

$$Y = \begin{cases} 1 & \text{if observation is from level 1 of B} \\ 0 & \text{otherwise} \end{cases}$$

then, $r^2_{xy} = \chi^2/N$. (See proof in Bishop, Fienberg, and Holland, 1975, p. 382).

Another related statistic is the phi (ϕ) coefficient developed by Karl Pearson. If a, b, c, and d denote cell frequencies as indicated by the table at the left, ϕ can be computed directly using the formula on the right.

		B: 1	B: 2
A	1	a	b
	2	c	d

$$\phi = \frac{bc - ad}{\sqrt{(a+b)\,(c+d)\,(a+c)\,(b+d)}}$$

But the formula for ϕ can be derived mathematically from the formula for the Pearson product-moment correlation coefficient. (See Glass and Stanley,

1970, pp. 158-160.) So we have

$$\phi^2 = r^2 = \chi^2/N.$$

Unlike the bivariate scatterplot of two continuously distributed variables as in Figure 1, the plot of X and Y in (1) does not show much. But the interpretation of r^2 (the coefficient of determination) as the proportion of variance in one variable explained by variation in the other holds for the categorical as well as the continuous case. (With a computer package like SAS (Barr, et al, 1976), it is easy to demonstrate this by calculating and printing predicted and residual scores, and computing their variances and σ_y^2.)

R x 2 Contingency Table

If the row variable has more than two categories, an R x 2 contingency table can be constructed, and the coding method in (1) can be extended. Code Y as before, and extend the coding of X as follows:

$$X_1 = \begin{cases} 1 & \text{if observation is from level 1 of A} \\ 0 & \text{otherwise} \end{cases} \qquad (2)$$

$$X_2 = \begin{cases} 1 & \text{if observation is from level 2 of A} \\ 0 & \text{otherwise} \end{cases}$$

. . . . . .

$$X_r = \begin{cases} 1 & \text{if observation is from level R-1 of A} \\ 0 & \text{otherwise} \end{cases}$$

Regression Y on $X_1, X_2, \ldots, X_r$ yields and R^2 which equals χ^2/N, where χ^2 is the test statistic for independence. Again, we can use our notions or R^2 to add meaning to the relationship between A and B tested using the value of χ^2.

A modification of the phi coefficient is made for contingency tables larger than 2 x 2: Cramer's V is given by

$$V = \left\{ \frac{\phi^2}{\min \{(R-1), (C-1)\}} \right\}^{\frac{1}{2}}.$$

(The denominator is the maximum that V attains, so that V ranges from 0 when no relationship is presented to a value of 1.) Substituting R^2 for ϕ^2 and solving for R^2 gives

$$R^2 = V^2 \times \min \{(R-1), (C-1)\} \qquad (3)$$

So from Cramer's V or the χ^2 value, we can compute a proportion of variance accounted for by the other.

R x C Contingency Table

The most general form of the contingency table has R rows for variable A and C columns for variable B. Variable A may be coded X_i as in (2). But, the coding of Y needs to be an orthogonal partition x of the variability in B. This is not difficult, using orthogonal polynominals, providing the frequencies in each level of A are equal. Assuming variable B to have four levels, code Y as follows:

$$Y_1 = \begin{cases} 1 & \text{if observation is from level 1 of B} \\ -1 & \text{if observation is from level 2 of B} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_2 = \begin{cases} 1 & \text{if observation is from level 1 or 2 of B} \\ -2 & \text{if observation is from level 3 of B} \\ 0 & \text{otherwise} \end{cases}$$

$$Y_3 = \begin{cases} 1 & \text{if observation is from levels 1, 2, or 3 of B} \\ -3 & \text{otherwise} \end{cases}$$

Then regress each Y_i ($i = 1, 2, \ldots, C-1$) on $X_1, X_2, \ldots, X_r$, denoting the respective multiple correlation coefficients by R_i.

Then

$$R^2 = \sum_{i=1}^{C-1} R_i^2$$

is equal to χ^2/N from the R x C table. Also, Cramer's V computed on the table using equaltion (3) yields and R^2 equal to χ^2/N.

It was hoped to related R^2 to Wilk's Lambda since multiple regression and multivariate analysis of variance have so much in common. (Kerlinger and Pedhazur, 1973, pp. 353 ff). But while apparently related, the exact connection could not be found by this author.

Conclusion

The purpose of this paper was to relate common statistics from contingency table analysis to the more familiar R^2 terminology in order to better understand the strength of the relation implied. The method of coding contingency tables in order to compute R^2's was shown and how R^2 relates to ϕ, V, and χ^2. It is not implied that all contingency tables be recoded so that multiple regression can be performed, but it is hoped that proportion of variance interpretations be done in addition to tests of significance.

References

Barr, A. J., Goodnight, J. H., Sall, J. P., & Helwig, J. T. A User's Guide to SAS 76. Raleight, NC: SAS Institute, Inc.

Bishop, Y. M. M., Fienberg, S. D., & Holland, P. W. Discrete Multivariate Analysis: Theory and Practice. Cambridge: MIT Press, 1975.

Glass, G. V, & Stanley, J. C. Statistical Methods in Education and Psychology. Englewood Cliffs, NJ: Prentice-Hall, Inc., 1970.

Kerlinger, F. N., & Pedhazur, E. J. Multiple Regression in Behavioral Research. New York: Holt, Rinehart and Winston, 1973.

McNeil, K. A., Kelly, F. J., & McNeil, J. T. Testing Research Hypotheses Using Multiple Linear Regression. Carbondale, IL: Southern Illinois University Press, 1975.